Article https://dx.doi.org/10.4314/iijikm.v9i2.1

Information Impact

Information Impact:
Journal of Information and Knowledge Management
2018, Vol. 9 (2) Pg 1 - 15
ISSN: 2141 – 4297 (print)
ISSN: 2360 – 994X (e-version)

Educational Data Mining for Students' Academic Performance Analysis in Selected Ethiopian Universities

Alemu K. Tegegne
Bahirdar University, Bahirdar, Ethiopia

Tamir A. Alemu
Bahirdar University, Bahirdar, Ethiopia

Abstract

Universities are working in a very dynamic and powerfully viable environment today. Due to the advent of information technology, they gather large volumes of data related to their students in electronic form in various formats like records, files, documents, images, sound, videos, scientific data and many new data formats. This study focuses on predicting performance of student at an early stage of the degree program, in order to help the university not only to focus more on bright students but also to initially identify students with low academic achievement and find ways to support them. The knowledge is hidden among the educational data set and it is extractable through data mining techniques. The aim of this paper is to use data mining methodologies to design and develop a Data Mining model to predict academic performance of students at the end of first year degree program in selected Ethiopian higher educational institutions (universities).The data of different undergraduate students has been mined with decision tree classifiers. A model is built using C4.5 Decision tree learning algorithm – generates five classification rule set classifiers (predictors) in an experiment. The experiment using a test data set produces 81.4% accuracy.

Keywords: **Educational Data, Educational Data mining, Decision tree, Classification rule, C4.5**

Introduction

The advances in the data mining field make it possible to mine the educational data and find information that allow for innovative ways of supporting teachers, students, and decision makers. The main functions of data mining are applying various methods and algorithms in different applications such as biological data analysis, financial data analysis, transportation, and forecasting/prediction (Mannilla, 1996). Such data mining and knowledge discovery applications have got a great attention due to its significance in decision making and it has become an essential component in various organizations including universities where educational data is mostly available. Moreover, knowledge extraction has got an additional opportunity since data mining techniques have been introduced into new fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities, etc.

The major motivation behind educational data mining in universities is that there are often information "hidden" in the data that are not readily evident, which may take enormous human effort,

time and therefore cost to extract. Furthermore, with exponential increase in the processing power of machines now available today, it is possible for data mining search algorithms to quickly filter data, extracting significant and embedded information as required. The ability to extract important embedded information in data suffices in many situations, helping organizations, companies, and research analysts make significant progress on different problems and decisions that are based on more information. By this task we extract knowledge that describes students' academic performance (achievement) up on the completion of the first year program. It helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising/counseling or devising appropriate teaching methods.

However, evaluating students' academic performance is generally a complex task since there are a number of factors that contribute to the success or failure of students in a course (Pidesky, Shapiro & Smyth, 1996). This study will explore the ability to mine available data using data mining algorithms, and therefore the information acquired be able to associate what factors are more predictive for the success or failure of students, how those critical factors can be fine-tuned for effecting better performance of students, and looking for optimal rule (model) in the predication of the overall students' academic performance. The main objective of this paper is to use data mining techniques and methodologies to design and develop a predictive model at the end of first year degree program to help for decision makers in a related domain.

Thus, this study is aimed to answer the following basic research questions

- To what extent data mining techniques applicable in predicting students academic performance through the use of academic (educational) factors?
- What variables or combination of variables collected can be used as predictors of students' performance at the desired academic level?
- Can we elicit the basic classification/association rule in analysing students' academic performance at the early stage of a program?
- How the discovered knowledge from academic data can aid decision makers to improve decision making processes?

Literature Review

One of the significant facts in higher learning institution (universities & colleges) is the explosive growth of educational data. They generate and gather large volumes of data with reference to their students, teachers and researchers in electronic form. However, they are data rich but information poor which results in unreliable decision making. One of the biggest challenges is unable to transform

large volumes of data into knowledge effectively to improve the quality of managerial decisions and impart quality education in the area. Good prediction of student's success in higher learning institution is one way to reach the highest level of quality in higher education system as well as to make good managerial decisions based on prior academic characteristics (pre-university information, socio-economic or demographic features, and soon). Application of educational data mining techniques is aimed at developing the methods that discover knowledge from data and used to uncover hidden or unknown information that is not apparent, but potentially useful (Han & Kamber, 2000).

The discovered knowledge can be used to better understand students' behaviour, to assist instructors, to improve teaching, to evaluate and improve teaching-learning system, to improve student academic performance; to improve curriculums and many others benefits. On the other hand, mining these educational data is used to find information that allow for innovative ways of supporting teachers, students and decision makers. Moreover, it increases the hope that the possibility to predict students' performance from such complex relationships can help facilitate the fine-tuning of academic systems and policies implemented in learning environments. As noted by Hijazi and Naqvi (2006) the application of educational data mining techniques in higher education institutions can be used to enhance learning, analyzing students' enrolment data to prevent drop-off and improve retention, to predict student retention at an early stage, and help to analyze the usage of learning materials given to students.

A number of works have investigated predicting performance at a university degree level. The study by Sembirin et al (2011) determines the relationship between students' demographic attributes, qualification on entry, aptitude test scores, performance in first year courses and their overall performance in the program. The investigation by Asif, Merceron and Pathan (2015) finds that performance in the first year of computer science courses is a determining factor in predicting students 'academic performance at the conclusion of the degree. A similar work by (Olani, 2008) employs the data mining technique random forests, essentially a set of decision trees, to predict students' graduate level performance (Master of Science, M.Sc.) by using undergraduate achievements (Bachelor of Science, B.Sc.). Another work in [8] predicts academic performance considering the data of two different universities. In the first case study, they use the data of undergraduate students of the University (CTU) in Vietnam to predict GPA at the end of 3rd year of their studies by using the students 'records (e.g. English skill, field of study, faculty, gender, age, family, job, religion, etc.) and 2nd year GPA. In the second case study, they consider the data of masters 'students of Asian Institute of Technology (AIT). By using students' admission information (like academic institute, entry GPA,

English proficiency, marital status, Gross National Income, age, gender, TOEFL score etc.) they predict the GPA of students at the end of 1st year of the master degree.

The study by Alnasir and Jaradat (2011) predicts students' university performance by using students' personal and pre-university characteristics. They take the data of 10330 students of a Bulgarian educational sector, each student being described by 20 attributes (e.g., gender, birth year and place, place of living, and country, place and total score from previous education, current semester, total university score, etc.). They have applied different data mining algorithms such as the decision tree C4.5, Naive Bayes, Bayesian networks, K-nearest neighbours (KNN) and rule learner's algorithms to classify the students into 5 classes i.e. Excellent, Very Good, Good, Average or Bad. The best accuracy obtained by all these classifiers is 66.3%.

Data Mining Definitions and Techniques

Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process -the sequences of steps identified in extracting knowledge from data.

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases (Lakshim, Krishna & Pumar, 2013) These techniques and methods in data mining are highlighted as follows:

***A*. Classification**

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large (Kumar & Saurabh, 2013). This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

B. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes (Bhanuprakash, Nijagunarya & Jayaran, 2011, Algur, Bhat & Kulkami, 2016). Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as pre-processing approach for attribute subset selection and classification.

C. Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

D. Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it (Bhanuprakash, Nijagunarya & Jayaran, 2017). During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

E. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to

generate rules with confidence values less than one (Mashat, Fouad, Yu & Gharip, 2013). However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

F. Decision Trees

Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. The three widely used decision tree learning algorithms are: ID3, ASSISTANT and C4.5 (Mirada et al, 2013). C4.5 is used as an implementation algorithm in this study.

Methodology

The data used in this study is collected from five selected regional universities in Ethiopia, namely Bahirdar university, Wollo University, Gondar University, D/Markos University and Debre Berhan University. The collected data comprises of quantitative aspects of a student in academic life properties from a year 2007 to 2012 excluding their name, ID, other indicative properties of a particular student. The data is purposely analyzed and a four year program of a student's data is used in the study. Though, the data is collected from where there is little differences in establishment, location, newness/oldness or other characteristics, we didn't differentiate among the data based on the assumption that all universities in Ethiopia has followed the same academic calendar, curriculum, seasoning, and mostly share the same academic characteristics. However, a data of which completely different academic characteristics with others is filtered and removed during the analysis stage.

As shown in figure 1, the collected data is analyzed and pre-processed using python 3.0. Python 3.0 is a power full text processing tool which all of the research team is familiar. It is used also to prepare the data to the format used in WekaV.0 decision tree classification model, which is the implementation tool.

The prepared data is classified in to 70/30 principle, i.e. 70 % of the data is used for training the model, and remaining 30% of the data set is used for testing the model. Then the accuracy of the model is measured how much of the test document is correctly predicted/ classified by the model.

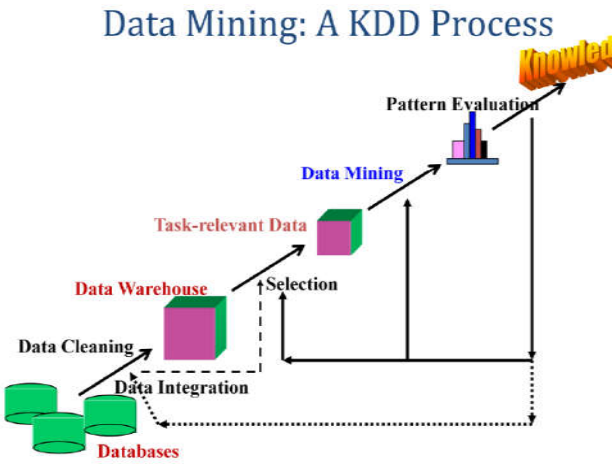

Fig. 1: Data mining: A KDD Process

Data Sets

The collected data sets mostly comprise dependent and/or independent academic variables

A. Independent variables

Such variables (predictors) were coded as:

PSGPA = Preparatory School Grade Point Average result

EUEE = Ethiopian University Entrance Examination result

FCI = Field Choice Interest

FYFSA = First Year First Semester Academic Achievement

B. Dependent (criterion) variables

Commutative Grade Point Avarage(CGPA) earned by each student at the end of first year second semester was the dependent variable of the study. It is annotated as

FYSSA = First Year Second Semester Academic Achievement

Data Mining Process

The data mining process is undertaken by evaluating Student related variables as shown in table 1.

Table 1: Student related variables

Variables	Descriptions	Possible values
PSGPA	Preparatory School Grade Point Average result	{High – >85%, Medium – >65% and Low>50%}
EUEE	Ethiopian University Entrance	{High – >600%, Medium – >450% and

	Examination result	Low>320%}
FCI	Field Choice Interest	{Yes, No}
FYFSA	First Year First Semester Academic Achievement	{Promoted – >2.00 Warning – >1.50 Sup >1.45}
FYSSA	First Year Second Semester Academic Achievement	{Promoted– >=2.00 Dissmisal – <2.00}

The domain values for some of the variables were defined for the present investigation as follows:

PSGPA: the cumulative, average Preparatory School grade Point Average result. It is split into three class values: High – average marks scored including and above 85 to100%, Medium – average marks scored including and above 65 to 85% and Low – average marks scored between and above 50 to 65%.

EUEE: average Ethiopian University Entrance Examination result of 700 mark in both natural and social science. It is categorized into three class sets as: High – average marks scored including and above 600 up to 700, Medium – average marks scored including and above 450 up to 600 and Low – the minimum passing (university entrance result, i.e. 320 in 2014/2015) up to 450.

FCI: Field Choice Interest that determines whether the field he/she has engaged is as per his interest/choice o not. Thus it has been classified in to two class values: Yes – the field is as per his interest, No –the field is not as per his interest

FYFSA: First Year First Semester Academic Achievement result obtained by the student. It has been classified as Promoted – students score (first semester GPA) including and above 2.00 ; Warning – students score(first semester GPA) including and above 1.50 ; and students score(first semester GPA) including and above 1.45}

FYSSA: First Year Second Semester Academic Achievement result obtained by the student and declared as dependent/response variable that determines the students' academic achievement in the university at the end of first year second semester. It has only to class sets: Promoted– >Average End semester Cumulative GPA is including and above 2.00 and Dismissal – Average End semester Cumulative GPA is below 2.00.

Findings and Discussion

The data set of 5729 students' record in this study was obtained from five university's 2014/2015 entry student records found in Amhara Regional state, Ethiopia. Table 2 shows the sample data set for the first 60 students of 2014/15 entry according the previous pre-defined academic variables as stated in table 2 below.

Table 2: Sample result data set

S.N	PSGPA	EUEE	FCI	FYFSA	FYSSA
1	High	Med	Yes	Pro	Pro
2	High	High	Yes	Pro	Pro
3	High	High	Yes	Pro	Pro
4	High	Med	No	Sup	Dis
5	High	Low	Yes	War	Dis
6	High	High	Yes	Pro	Pro
7	High	High	No	War	Pro
8	High	Low	No	Sup	Dis
9	High	Med	Yes	Pro	Pro
10	High	High	No	Pro	Pro
11	High	Med	Yes	Pro	Pro
12	High	High	Yes	Pro	Pro
13	High	Med	No	War	Pro
14	High	Low	No	Sup	Dis
15	High	Med	Yes	Pro	Pro
16	High	High	Yes	Pro	Pro
17	High	Med	Yes	Wa	Pro
18	High	Low	No	Sup	Dis
19	High	Med	Yes	Pro	Pro
20	High	High	Yes	Pro	Pro
21	Med	low	No	War	Dis
22	Med	High	Yes	Pro	Pro

23	Med	High	Yes	Pro	Pro
24	Med	Med	No	Sup	Dis
25	Med	Low	No	War	Dis
26	Med	High	Yes	Sup	Pro
27	Med	High	No	War	Pro
28	Med	Low	No	Sup	Dis
29	Med	Med	Yes	Pro	Dis
30	Med	High	No	Pro	Pro
31	Med	Med	Yes	Pro	Pro
32	Med	High	Yes	Pro	Pro
33	Med	Med	No	War	Dis
34	Med	Low	No	Pro	Pro
35	Med	Med	Yes	Pro	Pro
36	Med	High	Yes	Pro	Pro
37	Med	Med	No	War	Dis
38	Med	Low	No	Sup	Dis
39	Med	Med	Yes	Pro	Pro
40	Med	High	Yes	Pro	Pro
41	Med	Low	No	Sup	Dis
42	Med	Med	No	Sup	Pro
43	Med	Low	Yes	Pro	Dis
44	Med	Low	No	Sup	Dis
45	Med	Med	Yes	War	Pro

46	Med	High	Yes	Pro	Pro
47	Med	Low	No	War	Dis
48	Med	Med	No	Sup	Pro
49	Med	Low	No	War	Dis
50	Med	Low	No	War	Dis
51	Low	Med	Yes	War	Pro
52	Low	High	Yes	Sup	Pro
53	Low	Low	No	War	Dis
54	Low	Low	No	Sup	Dis
55	Low	Med	Yes	Pro	Pro
56	Low	High	No	Sup	Pro
57	Low	Low	No	War	Dis
58	Low	Low	No	War	Dis
59	Low	Med	Yes	Sup	Dis
60	Low	High	Yes	Pro	Pro

*Note that – Med is for medium, Sup is for supplementary exam, Pro is for promoted, War is for warning, and Dis is for Dismissal.

To work out the information gain for A relative to S, we first need to calculate the entropy of S. Here S is a set of 60 examples are 20 “High”, 30 “medium”, 10 “low” .

Entrophy (S) = – PHigh Log 2(PHigh) –Pmedium Log2(Pmedium)– PLowLog2(P Low)

$= - \left(\frac{20}{60}\right)\text{Log 2}\left(\frac{20}{60}\right) - \left(\frac{30}{60}\right)\text{Log 2}\left(\frac{3}{60}\right) - \left(\frac{10}{60}\right)\text{Log 2}\left(\frac{10}{60}\right)$

= 1.946

To determine the best attribute for a particular node in the tree we use the measure called Information Gain. The information gain, Gain (S, A) of an attribute A, relative to a collection of examples S,

Gain(S PSGPA) = Entropy(S) – $\frac{|S\text{high}|}{|S|}$ Entropy(SHigh)

$$- \frac{|Smedium|}{|S|} \text{Entropy(SMedium)}$$

$$- \frac{|low|}{|S|} \text{Entropy(SLow)}$$

Thus, it results the information Gain is presented as shown in table 3.

Table 3: Gain Values

Gain	Values
Gain(S, PSGPA)	0.567036
Gain(S, EUEE)	0.515125
Gain(S, FCI)	0.025410
Gain(S, FYFSA)	0.437022

PSGPA has the highest gain, therefore it is used as the root node as shown in figure 2.

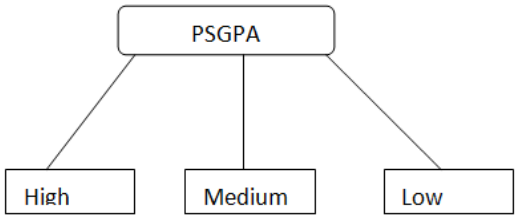

Fig. 2. PSGPA as root node

Gain Ratio can be used for attribute selection, before calculating Gain ratio Split Information is shown in table 4.

Table 4: Split Information

Split information	**Value**
Split(S,PSGPA)	1.37565
Split(S, EUEE)	1.43755
Split(S,FCI)	1.98664
Split(S,FYFSA)	1.50263

Gain Ratio is shown in table 5 bellow.

Table 5: Gain Ratio

Split information	Value
Gain ratio (S,PSGPA)	0.41535
Gain ratio (S, EUEE)	0.34565
Split(S,FCI)	0.02136
Gain ratio (S,FYFSA)	0.28874

This process goes on until all data classified perfectly or run out of attributes. The knowledge represented by decision tree can be extracted and represented in the form of IF-THEN rules.

Table 6: Rule Set generated by Decision Tree

IF PSGPA =’High’ AND EUEE =’High’ OR ‘medium’ AND FYFSA=’promoted’ THEN FYSSA=’Promoted’
IF PSGPA =’High’ AND EUEE =’High’ AND FYFSA =’Promoted’ OR ‘warning’ then FYSSA=’Promoted’
If PSGPA =medium AND EUEE =’High’ OR ‘medium’ AND FCI=’yes’ THEN FYSSA=’Promoted’
IF PSGPA =’low’ AND EUEE=’low’AND FYFSA=’sup’ THEN FYSSA=’Dismissal’
IF PSGPA =’low’ AND EUEE=’low’ AND FYFSA=’sup’ OR ‘warning’ THEN PSGPA=’low’

One classification rules can be generated for each path from each terminal node to root node. Pruning technique was executed by removing nodes with less than desired number of objects. IF-THEN rules may be easier to understand as shown in table 6. In this experiment, the decision tree classifier produces 81.4% accuracy with a new test data set.

Conclusion

This study focused on predicting students’ academic performance through educational data mining at an early stage of the degree program, in order to help the university and other decision makers not only to focus more on bright students but also to initially identify students with low academic achievement and find ways to support them. This paper is aimed to use data mining methodologies to design and develop a Data Mining model to predict academic performance of students at the end of first year degree program. The data of different undergraduate students has been mined with different classifiers. In this study, the dependent and/or independent variables have been

identified and utilized. These variables are used in an experiment using decision tree. Hence, classification rule is generated and applied on a new test data set and result in significant accuracy. The findings of the study have important implications for educators, teachers, counselors, university curriculum designers, students and other decision makers.

References

Algur,P.S., Bhat,P., & Kulkarni, P.(2016). Educational Data Mining: Classification Techniques for Recruitment Analysis. *IJMECS* Vol.8, No. 2,pp 59-65.

Asif, R., Merceron, A.,& Pathan, M. K (2015). Predicting Student Academic Performance at Degree Level: A Case Study*, I.J. Intelligent Systems and Applications*, 01, 49-61

Bhanuprakash,C., Nijagunarya, Y. S., & Jayaram, M.A(2017). Clustering of Faculty by Evaluating their Appraisal Performance by using Feed Forward Neural Network Approach. . *IJISA* Vol. 9, No. 3, pp.34-40.

Han, J., & Kamber, M.(2000). *Data Mining: Concepts and Techniques*, Morgan Kaufmann

Hijazi, S.T.,& Naqvi, R.S.M (2006). Factors affecting student‟s performance: A Case of Private Colleges", *Bangladesh e-Journal of Sociology*, Vol. 3, No. 1,

Kumar, A. & Saurabh P. (2013). Classification Model of Prediction for Placement of Students. *IJMECS* Vol.5, No. 11.pp.49-56.

Lakshmi ,K.R., Veera, M., Krishna, S. & Kumar, P. (2013).Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability. *IJMECS* Vol.5, No. 8, pp. 8-17.

Mannila,H (1996). Data mining: machine learning, statistics, and databases, IEEE.

Mashat, A.F., Fouad, M.M., Yu, P.S., & Gharib, T.F(2013). Discovery of Association Rules from University Admission System Data.. *IJMECS* Vol.5, No. 4,pp.1-7.

T.Miranda, T., Lakshmi, A., Mumtaj, M.R. Begum,V.,& Venkatesan. P (2013). An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data. *IJMECS* Vol.5, No. 5, pp.18-27.

Olani, A**.** (2008). Predicting first year university students' academic success. *Netherlands Journal of Research in Educational Psychology*, 7(3), 1053-1072.

Piatesky,U. F., Shapiro, G., & Smyth, ,P. (1996). *From data mining to knowledge discovery in databases,* AAAI Press / The MIT Press, Massachusetts Institute Of Technology

Sembiring,S et al. (2011). Prediction of student academic performance by an application of data mining techniques. 2011 International Conference on Management and Artificial Intelligence, IACSIT Press, Bali, Indonesia.